技术交底书

1 背景技术

第二代测序技术的发展使生命科学快速进入了基因组时代。对自身遗传密码的掌握和破译，使人类在生物医学的研究中拥有了前所未有的机会。越来越多的证据表明，来自于基因组的变异是造成人类疾病产生的主要原因，且某些突变大量存在：如单核苷酸多态性位点（Single Nucleotide Polymorphism，SNP），因其数量多、分布广泛，适于快速和大规模筛查等特点，已成为第三代遗传标记。SNP 主要是指在基因组水平上由单个核苷酸的变异所引起的 DNA 序列多态性。它是人类可遗传的变异中最常见的一种。占所有已知多态性的 90%以上。除此之外，还有 CNV(copy number variation), InDel (insertion-deletion)等各类基因组变异。尽管这些突变可能只涉及到单个碱基的变异，其后果却可影响到基因、蛋白及性状的直接改变。不仅单个突变可直接导致遗传疾病的产生，研究表明许多复杂疾病也是由多个突变共同作用的结果。目前，基因组突变解析已成为基因组研究的重要内容，并在疾病诊断与筛查、复杂性状和疾病的遗传解剖、不同人群遗传特征研究中得到了显著应用。由于基因组变异及其导致的氨基酸突变不仅影响人类疾病的发生发展，还影响了机体对病原体、药物和疫苗等的反应，故也是个性化医疗的关键。如对 SNP 的研究被认为是人类基因组计划走向应用的重要步骤。

由于基因突变在基因组中大量存在，每个个体基因组中都存在着相当一部分未知突变，故已知突变数量随着个体基因组的解析急剧增长、不断更新。为更好的管理、理解和使用基因突变，尤其是 SNP 信息，有关数据库应运而生。例如由 NCBI 管理和维护的 dbSNP(1)，ClinVar(2)，以及 HGMD(3)和 OMIM(4)。其中，dbSNP 侧重于收集各类已知 SNP 信息，HGMD 专注于可遗传的基因变异，OMIM 作为最初记录人类疾病表型信息的数据库，只囊括已有确切证明与遗传疾病相关的突变信息。ClinVar 作为上述数据库信息的整合，其信息来源的时效性有限，由于需通过人工验证，检索结果与最新发表文献数据可能存在时间差。从应用角度来说，与疾病相关的突变具有最重要的研究价值，需要一个来自前沿、信息全面，可快速更新的数据库。同时，由于突变信息的数据量大，更新速度快，人工

收集不仅成本高、速度慢，还存在信息遗漏和可信度低等问题。在这两方面，上述数据库的应用仍显不足。且由于部分数据库存在收费等问题，限制了用户对最新数据的获得。

为解决上述问题，我们拟从原始文献出发，利用文本挖掘技术提取突变信息及与疾病的相关关系，建立基于研究前沿、数据全面可靠、可快速更新的人类遗传疾病相关 SNP 数据库，提供相关疾病、基因、突变和文献等的详细信息。本发明将为人类遗传疾病的研究提供重要信息数据库，也是国内首个基于文本挖掘技术的基因突变数据库，将极大提升我国基因组学研究的基础实力。

2 技术方案的详细叙述

本发明从原始文献出发，利用网络搜索和文本挖掘算法技术提取突变信息及与疾病的相关关系，建立基于研究前沿、数据全面可靠、可快速更新的人类遗传疾病相关 SNP 数据库，提供相关疾病、基因、突变和文献等的详细信息。见图 1，图 2。

2.1 获取文献摘要

根据疾病列表从 NCBI 中搜索相关文献摘要。

2.2 前处理（获取摘要标题、内容、影响因子等）

对每一篇文献摘要进行分解，获取其发表杂志机构、发表日期、文献标题、作者、摘要内容、PMID。

2.3 获取基因、突变信息并进行分类过滤处理并注释

根据基因列表和突变正则表达式语义库，对文献摘要内容进行基因和突变信息的提取和分类。（分 DNA 突变、蛋白质突变（氨基酸突变）、Rs 号三类）。并对这三类突变分别比对 hg19 文档进行过滤和注释。

2.4 后处理（基因突变得分判定）

对于某个疾病，用统计分析的方法构建对应于该疾病的相关词语义库，利用该相关词语义库给每一篇文献摘要打分，即文献摘要分数。对于基因，根据统计分析的方法有相应的基因相关度分数。综合两者得分，即为基因突变的得分判定过程。

2.5 构建金标准数据集，对数据进行验证

根据人工构建的金标准数据集，对系统获取的数据进行性能评估。共三个指标：准确率（Precision）、召回率（Recall）、F1 值（F1-Measure）。

3 有益效果

本发明以人类遗传病为出发点，在已有科研文献的基础上，通过文本挖掘算法构建多个语义库，使用机器学习和挖掘海量文献内容，收集整理突变信息。经人工验证构建金标准数据库，使用科学的计算公式验证结果的准确性。其有益之处在于：

(1) 网络爬虫和文本挖掘技术的应用保证了海量数据处理和及时快速更新。

(2) 多种验证方式保证了所得数据的全面、准确及可信度。

(3) 多个语义库的构建不仅保证本项目的顺利实施，也将为类似研究工作提供参考和基础。

4 具体实施例

以疾病 Asthma 为例：

4.1 以疾病名（例：Asthma）为搜索词，利用 NCBI 提供的 Pubmed 网络搜索接口获取关于 Asthma 的所有相关文献摘要资源。

4.2 对每一篇文献摘要进行分解，获取其发表杂志机构、发表日期、文献标题、作者、摘要内容、PMID。（由于文献摘要的格式多样化，因此需要根据不同的文献摘要格式制定相应的分解策略。）

获取文献摘要内容后，根据 Asthma 的同义词对文献摘要进行过滤，若该篇文献摘要内容包含 Asthma 的同义词（该同义词列表来源于 Comparative Toxicogenomics Database [1]），则保留。Asthma 的同义词如下：

(Asthma, Bronchial|ASTHMA, DIMINISHED RESPONSE TO ANTILEUKOTRIENE TREATMENT IN, INCLUDED|ASTHMA-RELATED TRAITS, SUSCEPTIBILITY TO ASTHMA,PROTECTIONAGAINST,NCLUDED|Asthmas|ASTHMA,SUSCEPTIBILITY TO|Bronchial Asthma)

根据杂志机构影响因子列表（2013 年 9 月的影响因子列表）对该篇文献摘要进行影响因子注释。（由于文献摘要中杂志机构的表达方式多样化，因此需要根据不同的表达方式制定相应的注释策略。）

4.3 根据 Comparative Toxicogenomics Database [1]中关于 Asthma 的基因列表和突变正则表达式语义库，对 Asthma 相关的文献摘要内容进行基因和突变信息的提取和分类。（分 DNA 突变、蛋白质突变、Rs 号三类）。并对这三类突变分别比对 hg19 文档进行过滤和注释。详细如下：

对于 DNA 突变:（例如 BRCA1　c.123A>T）若在 hg19 中的转录本中 BRCA1 基因的第 123 号位置为 A 碱基，则保留。以此类推。并对保留的 DNA 突变进行染色体号、坐标位置、rs 号、+/-链、周边详细突变信息的注释。

对于蛋白质突变:（例如 BRCA1　p.K123L）若在 hg19 中的转录本 BRCA1 基因的第 123 号密码子翻译为 K，则保留。以此类推。由于无法判断是密码子中的哪一个碱基发生突变，因此不对其进行注释。

对于 Rs 号，（例如 rs12345678）若在 hg19 的 dbSNP137 库中比对上基因，则为其注释上基因。若比对不上基因，则基因置为 NULL。以此类推。（由于 Rs 号可对应 0,1,2 个基因，并且不能根据 hg19 的转录本进行过滤，因此默认为所有 Rs 号都是正确的。）并对 Rs 号进行染色体号、坐标位置、+/-链、周边详细突变信息的注释。

4.4 对于 Asthma 疾病，以所有与 Asthma 相关的文献摘要为训练数据，采用统计分析的方法构建对应于 Asthma 的相关词语义库。

具体过程如下：

（1） 窗口：即该篇文献摘要的单词数。（以 Asthma 主词，R 为相关词为例。）

（2） 共现次数：以每一篇文献摘要为窗口，以 Asthma 为主词，计算该篇文献摘要中 R 与 Asthma 在该篇文献摘要中的共现次数 Part_Co_occurrence、并计算该篇文献摘要中 R 与 Asthma 在所有文献摘要中的共现次数 All_Co_occurrence。

（3） 平均距离：以每一篇文献摘要为窗口，以 Asthma 为主词，先计算得出 Asthma 的位置，若该篇文献摘要中不存在 Asthma，则将 Asthma 的位置取为窗口的一半。计算该篇文献摘要中 R 与 Asthma 的最小距离，若该篇文献

摘要 R 出现多次，取其他们的平均距离 d。如果 Asthma 的相关词语空间中没有 R， 则把 R 加入 Asthma 的相关词语空间，并将 R 的 D 值设置为 d；否则，将 R 的 D 值设置为 D=（Dc+d）/(c+1).其中 c 为 All_Co_occurrence。

（4） 信息熵：$$\eta(i)=-\sum_{j=1}^{t}[(a_{ij}/\sum_{j=1}^{t}a_{ij})\ln(a_{ij}/\sum_{j=1}^{t}a_{ij})]$$

式中：aij 为词语 i 在第 j 篇文本中出现的次数；t 表示文本总数。η(i)值越大，表明这个词的信息熵越大，即该词是噪音词语的可能性也越大。计算信息熵的过程中，信息熵的值可能为零，为避免将基因丢失，对于此类情况将其相关度置为 1。

（5） 相关度：$$r=(c\times\alpha)\div(\alpha+d)\div\eta$$

c 为共现次数（All_Co_occurrence），d 为平均距离（D），η 为信息熵。r 为某个相关词与主词的相关度；α 为一个距离系数，目的是为了相对于 c，降低 d 对 r 的影响。α 为平均窗口大小的 1/3。

（6） 最后，构建好的语义库格式为 R 18.88；即每一个相关词都有一个对应于 Asthma 主词的相关度分数。

基因突变得分规则：

（1） 根据 Comparative Toxicogenomics Database 中关于 Asthma 的基因列表，从构建好的语义库中提取基因部分，因此对应于 Asthma 的每一个基因都有一个相关度分数。并对基因相关度分数做了归一化处理，即将所有基因相关度分数分别除以最大的基因相关度分数并乘以 20，则最后得到的基因相关度分数最大阈值为 20。

（2） 提取去除基因部分后的相关词语义库中排名前 20 位的相关词作为判定文献摘要分数的相关词语义库。以 A 文献摘要为例，若 A 文献摘要包含 20 个相关词中的 18 个，则 A 文献摘要得分为 18，即文献摘要分数。则最后得到的文献摘要分数最大阈值为 20。

（3） 经过三种得分策略的尝试（侧重基因、侧重文献摘要、两者兼顾），发现兼顾两者得分的效果较好，突变得分分布较集中。因此采取基因得分最大值为 20，文献摘要得分最大值为 20 的两者兼顾得分策略。综合两者得分的公式（此公式参考 F1 值计算公式）如下：

2*Score(gene)* Score(abstract)/(Score(gene)+Score(abstract))

4.5 规律：得分在 9 到 10 分以上的基因突变对准确率为 80%左右。整理得到的最终数据，并进行人类疾病突变数据库网站的构建。见图 3，图 4，图 5。

5 附图及说明

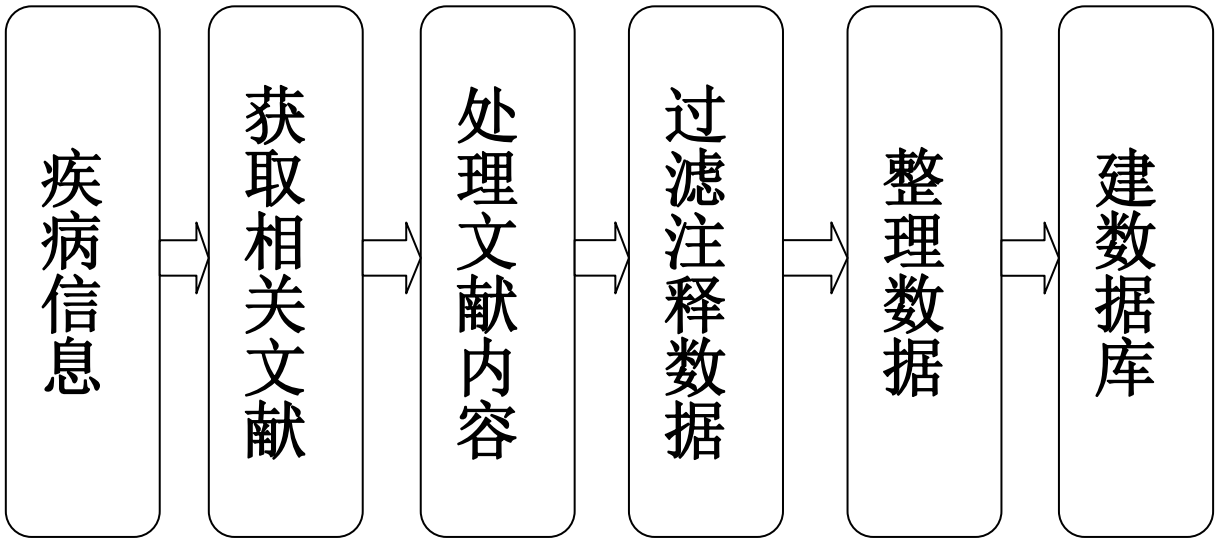

图 1 文本挖掘方法流程示意图

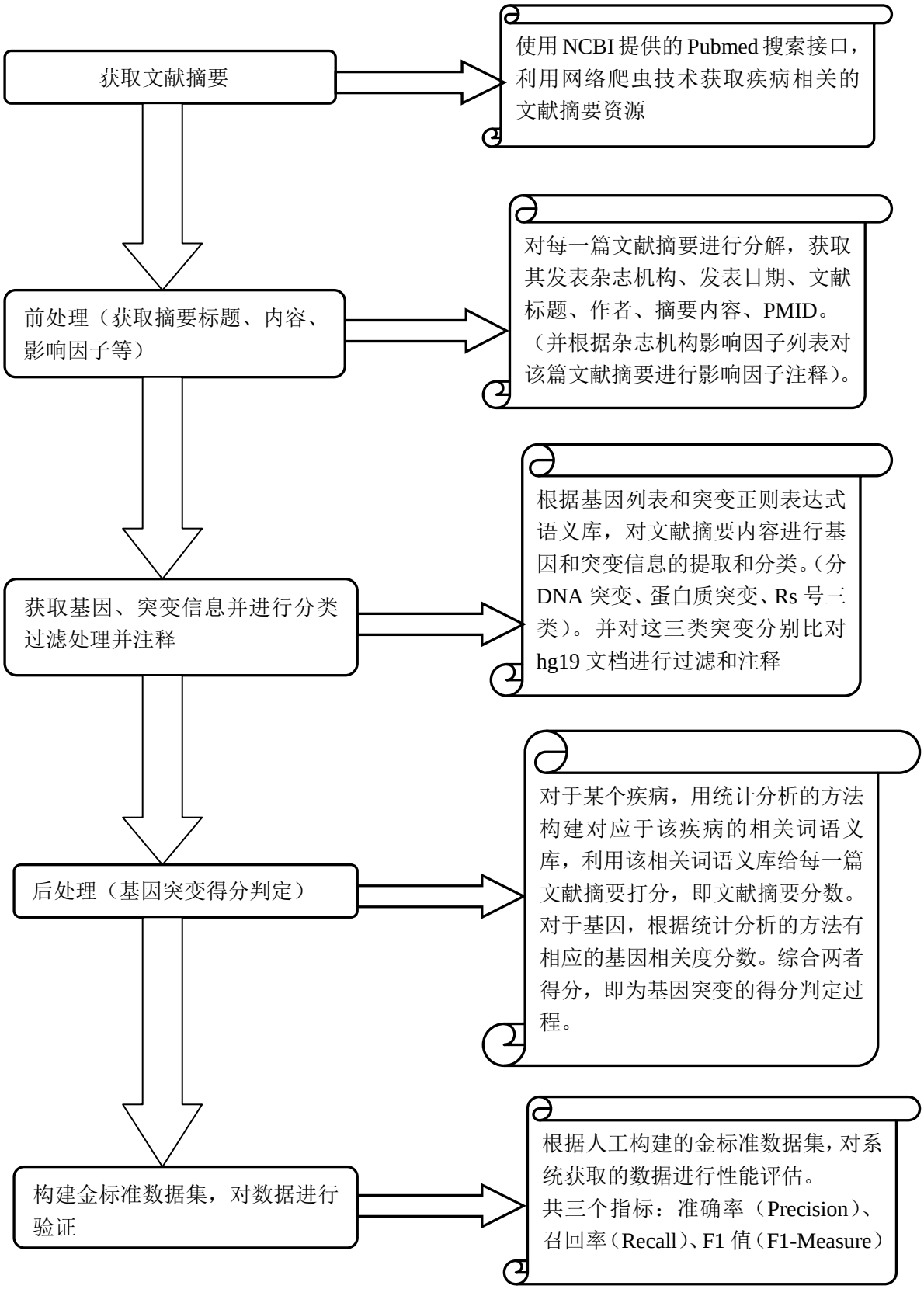

图 2 文本挖掘方法流程详细示意图

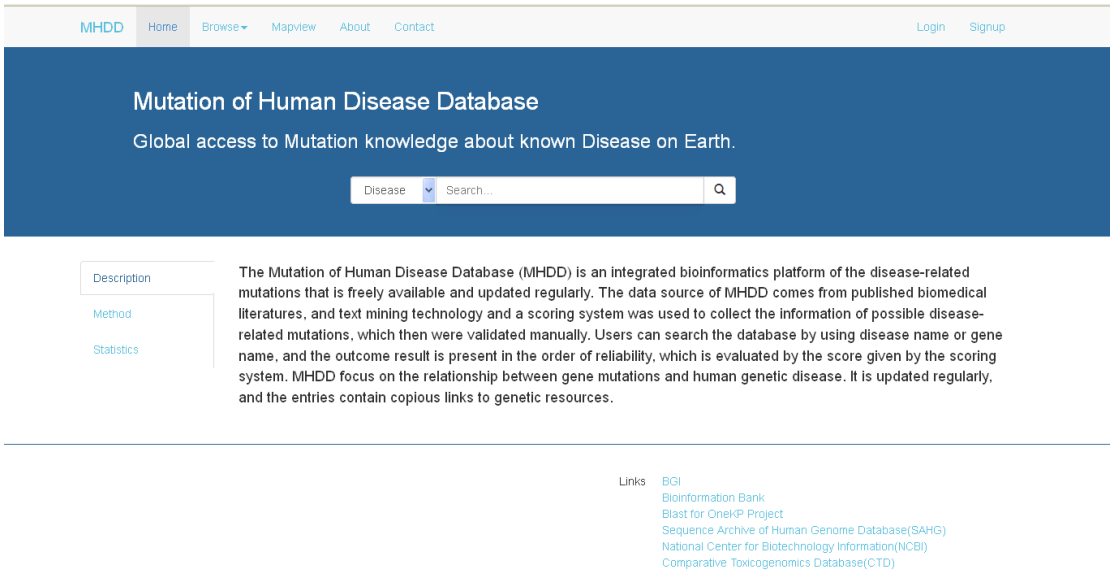

图 3 人类疾病突变数据库主页

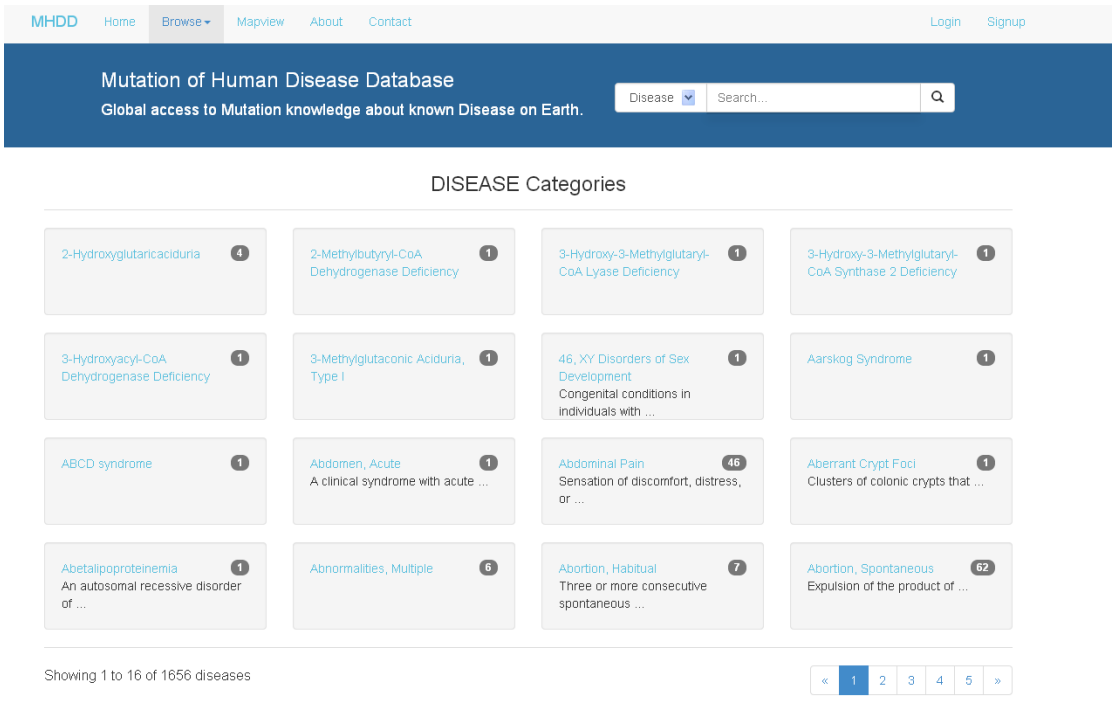

图 4 人类疾病突变数据库疾病搜索页

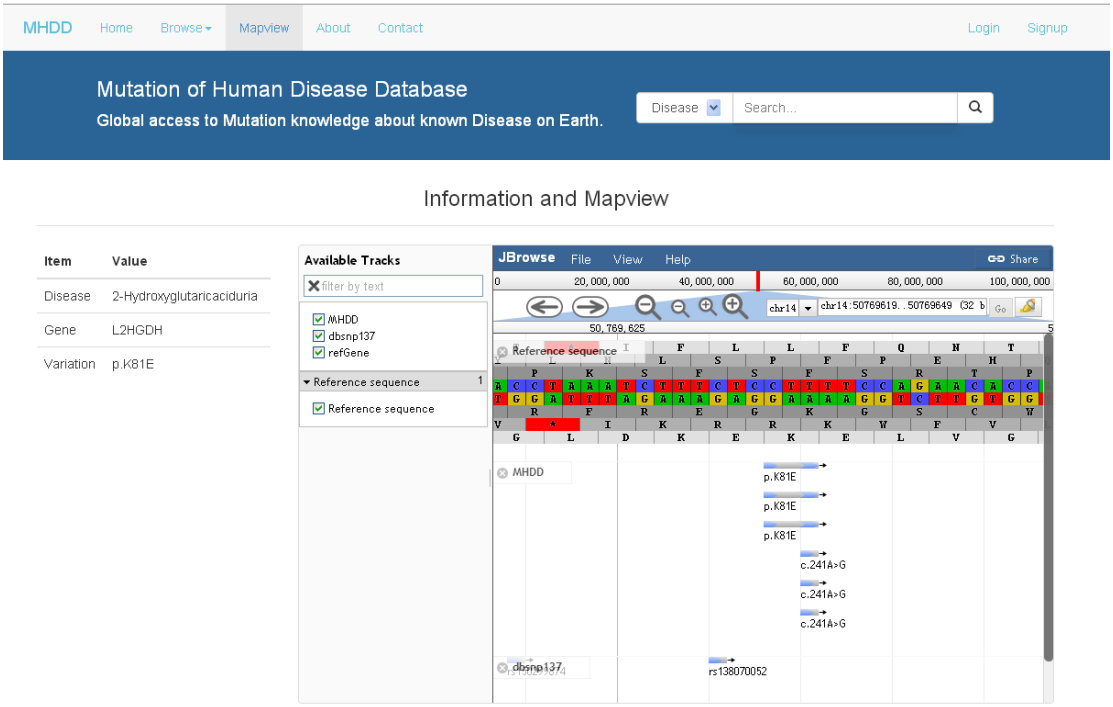

图 5 人类疾病突变数据库突变 Mapview 页面